EVA
Berlin

EVA Berlin 2017 Conference
8 - 10 November 2017
Kunstgewerbemuseum (Museum of Decorative Arts)
Kulturforum Potsdamer Platz, Matthäikirchplatz 8, 10785 Berlin

Digitale Zwillinge sollten sich nicht zu sehr ähneln und «getrennt wohnen»

Bernd Kulawik[a]

[a] Architektur- und Musikhistoriker; Bern, Schweiz, be_kul@me.com

KURZDARSTELLUNG: Neben den Gefahren, die Projekten in den *Digital Humanities* drohen, weil das Problem der wirklich *langfristigen* Sicherung und nachhaltigen Verfügbarkeit der Daten bisher nicht als zufriedenstellend gelöst angesehen werden kann und muss, gibt es selbst für die aktuellen «kurzfristigen» Lösungsansätze mit Verfügbarkeitsdauern – je nach Art der verwendeten Software und Datenformate – von 10, vielleicht 20 bis maximal 50 Jahren Gefahren: Obwohl des Stichwort «Digitale Zwillinge» im Call for Papers für diese EVA sich offensichtlich eher auf die Zwillingspaare Objekt–Digitalisat bezieht, soll im Folgenden skizziert werden, warum auch das viel ähnlichere Zwillingspaar aus Digitalisat und Digitalisat Gefahren ausgesetzt ist, die in jedem Projekt der *Digital Humanities* berücksichtigt werden sollten. Denn üblicherweise erfolgt die Datenhaltung und Speicherung «*langfristig*» weder verteilt noch betriebssystem- oder gar software-unabhängig, so dass physische oder digitale Zerstörungen bspw. durch erpresserische Verschlüsselung ein ernst zu nehmendes Problem darstellen.

1. EINLEITUNG

Nachdem ich hier in den Jahren 2016 [1] und 2017 [2] in Übereinstimmung mit und im Anschluss an die «Väter» des TCP/IP, Vinton Cerf (*Google*-Vizepräsident und langjähriger Präsidenten der ACM) und Robert Cahn das «Horrorszenario» eines «Digital Dark Age» und des Verschwindens aller unserer Daten im «Information Black Hole» in spätestens 50 Jahren beschrieben und die «maximalistische» Forderung für eine Lösung vorgeschlagen hatte, die einem vollständigen «Reboot» der digitalen Infrastruktur in einem *gezielt langlebigen* und *entwicklungsfähigen* System unter Verwaltung einer Institution mit «Ewigkeitsgarantie» entspricht, gibt mir das Thema des «Digitalen Zwillings» die Möglichkeit, einige Ideen für die aktuell üblichen, vor diesem Hintergrund allesamt als «kurzfristig» anzusehenden Zwischenlösungen zu präsentieren: Denn neben der Verwendung langfristig nicht überlebensfähiger Software birgt die häufig anzutreffende Beschränkung der jeweiligen Softwarelösung auf *ein* Projekt und seiner Daten auf *eine einzige* Installation und ggf. in lokaler Nähe aufbewahrte Backups ein weiteres, häufig vernachlässigtes Gefahrenpotential für die Daten, die in wissenschaftlichen und kulturellen Kontexten aufwendig und kostenintensiv erhoben, verarbeitet und präsentiert werden. Damit droht nicht nur die langfristige Vernichtung von Steuergeldern, sondern vor allem auch von Lebens- und Arbeitszeit in einem kaum vorstellbaren Ausmaß. Während man also leider immer noch mit einiger Sicherheit voraussagen können wird, dass kaum ein «Digitaler Zwilling» eines realen Objekts dieses selbst in seiner Lebensdauer übertreffen wird, sollte zumindest der Gefahr der physischen oder elektronischen Zerstörung dieser Zwillinge durch Schaffung «digitaler Zwillinge» begegnet werden, die sich weder «kennen» noch zu sehr ähneln.

VISIT…

LANZAROTE Caliente.COM

2. VORGESCHICHTE

Es sollte langsam im allgemeinen Bewusstsein von Forschenden, aber eigentlich von allen Computerbenutzern angekommen sein, dass ihre in den zurückliegenden Jahren oder gar Jahrzehnten erhobenen Daten der bisher nicht abgewendeten Gefahr ausgesetzt sind, mitsamt der zu ihrer Benutzung und Interpretation notwendigen Software – vom Anwendungsprogramm bis zum Betriebsstystem – einer Haltbarkeit von ca. 20, in wenigen, sehr einfachen Fällen wie den Formaten TXT und PDF/A vielleicht 50 Jahren zu unterliegen. [3] Davor warne nicht nur ich als in «Computerei» dilettierender Musik- und Architekturhistoriker, sondern seit geraumer Zeit auch der «Vater des Internet», Vice President von Google und langjährige Präsident der ACM (Association for Computing Machinery), der größten einschlägigen Ingenieursvereinigung, Vinton Cerf. [4] Zumindest bei *ihm* sollten Sie sicher sein, dass er weiss, wovon er spricht, und seine Warnungen es verdienen, Ernst genommen zu werden.

Eine breit anwendbare Lösung für dieses langfristige Problem ist aber m.W. genausowenig in Sicht wie überhaupt das anzumahnende Bewusstsein seitens der Historikerinnen und Historiker – was mich insofern wundert, als man sich fragen kann: Wer, wenn nicht *sie* bzw. wir sollten auf diese Frage der zukünftig «historischen» Verfügbarkeit ihrer (nicht nur) digitalen Objekte – ihrer Arbeitsgrundlagen! – und ihrer Forschungsergebnisse längst zumindest ein paar Gedanken «verschwendet» haben… und sich konsequent weigern, irgendwelche «modische» Software mit «innovativen» «bells & whistles» einzusetzen, deren Lebensdauer in Jahren – mitsamt der ihrer Daten – man an zwei Händen abzählen bzw. abschätzen kann? [5]

Aber neben diesem langfristigen Problem – das aus *historischer* Sicht natürlich ein extrem kurzfristiges ist – haben selbst die heute üblichen, *wirklich* kurzlebigen Projekte und die dabei eingesetzte Software einige Probleme, die im Folgenden skizziert werden sollen.

3. DIGITALE ZWILLINGE SOLLTEN SICH «GETRENNT WOHNEN»

Immer noch werden in der Forschung und in der Kommunikation über sie nicht nachhaltige Datenformate und Softwarelösungen verwendet – wie in diesem Word-Dokument, das Sie gerade lesen. Dabei ist die Kommunikation selbst bekanntlich ein wesentlicher Teil der Forschung, erst recht in Zeiten, in denen die einzelne Forscherin in Archiv und Bibliothek oder der einzelne Forscher am Schreibtisch selbst in den historischen und Geisteswissenschaften eine aussterbende Spezies ist: Denn selbst diese kommen natürlich ohne Computer, Datenbanken und Internet heute kaum zurecht.

Für letzteres wurde bekanntlich vor fast 30 Jahren durch Tim Berners-Lee das *World Wide Web* entwickelt, mit dem man bequemer und schneller über das Internet Ideen und Publikationen austauschen können sollte. Dass es schnell von einem *peer-to-peer*-Ansatz zu einer *Top-Down*-Lösung wurde, mag anfangs noch der geringen verfügbaren Leistungskraft von Computern geschuldet gewesen sein – in Zeiten aber, in denen jeder einen «Computer» namens Smartphone mit sich herumträgt, der über ein Vielfaches der Leistungsfähigkeit der größten damaligen Server verfügt, ist diese Ungleichheit der Benutzer bzgl. ihrer unterschiedliche Berechtigungen im Zugang zu bzw. in der Verfügbarkeit über Server-Leistungen nicht mehr zu rechtfertigen. [6]

Diese könnte bereits auch eine Teillösung des hier zu behandelnden Problems bieten, indem zumindest Publikationen und ihnen zugrundeliegende Daten durch häufiges Teilen dezentral gelagert und so vor Datenverlust an einem spezifischen Ort – bspw. dem ihrer Erzeugung – geschützt würden. Das ist ja bereits ein zentraler Gedanke hinter der Netzwerkstruktur des Arpanets, des Vorläufers des Internets, für welches Vint Cerf und Rob Kahn 1972 die Protokollfamilie TCP/IP entwickelten…

Da wir von solch einer Lösung quasi nach dem «Graswurzel-Prinzip» bzw. in einem Bottom-Up-Ansatz noch immer weit entfernt sind und sie für sehr große Datenmengen wie Bilddatenbanken mit zigtausenden Dateien nicht wirklich praktikabel erscheint, wäre über eine andere Lösung nachzudenken, die eine verteilte Datenhaltung für jene Daten sicherstellt, die heute aktuell in Forschungsprojekten erzeugt werden oder bereits vorhanden sind. Natürlich ist dabei ein Netzwerk aus verteilten Speicherstrukturen, wie es der *Rat für Informations-Infrastrukturen* für Forschungsdatenmanagement – mit «nur» ca. 25-jähriger Verspätung vorschlägt – als erste Grundvoraussetzung sehr zu begrüßen: Es kann m.E. aber nicht ausreichend sein. Denn eine solche *nationale* Infrastruktur ist immer noch auf einen vergleichsweise engen geographischen Raum beschränkt und daher weder vor Natur- noch technischen Katastrophen geschützt. Man muss ja nicht gleich an einen Tschernobyl-artigen Ausbruch in den AKWs von Tihange

oder Fessenheim oder den Ausbruch des Supervulkans unter den *Phlegräischen Feldern* und damit die schlagartig notwendige Evakuierung riesiger Landstriche denken: Es genügt m.E. schon, sich einen 3-4tägigen großflächigen Stromausfall mit allen seinen (sozialen) Konsequenzen auszumalen. Die Haltung: «Irgendjemand wird sich schon um unsere Daten kümmern, wenn wir es nicht mehr können», halte ich nicht nur diesbezüglich für naiv. Aber vielleicht sehe ich ja nur zu viele dystopische Filme?
Trotzdem sollte für unsere Forschungsdaten nach einer Lösung gesucht werden, mit welcher diese Daten mehrfach in sehr großer räumlicher Entfernung, möglichst auf anderen Kontinenten, regelmäßig gespeichert werden. Und natürlich sollte dies nicht nur für Backup-Dateien gelten, sondern für die *gesamte* zur Benutzung der Daten notwendige Software-Infrastruktur. Die Vorstellung, irgendjemand würde sich schon die Mühe machen, aus den Backups eine historisch-geisteswissenschaftliche Forschungsdatenbank zu rekonstruieren, wenn ihre Ersteller dies nicht mehr können, dürfte ebenfalls mindestens illusorisch sein…

4. DIGITALE ZWILLINGE SOLLTEN SICH NICHT «ÄHNELN»

Aber eine geographisch verteilte Datenhaltung kann noch nicht als ausreichend angesehen werden: Es ist jederzeit zu erwarten, dass eine noch unbekannte Sicherheitslücke in identischen Installationen einer Forschungsanwendung gezielt ausgenutzt werden kann. Von gravierenden Fehlern in den weitverbreiteten PHP-basierten Webdatenbank-Anwendungen hört man bspw. ja fast im Wochentakt. Aber ich vermute, alle Ihre entsprechende Software einsetzenden Projekte verfügen über mehrere IT-Spezialisten, die im Dreischichtsystem 24/7 bereit stehen, jederzeit solche Sicherheitslücken zu schließen – oder nicht?
Wenn lediglich die Backups einer Forschungsdatenbank verteilt gehalten werden, könnte man meinen, einem solchen Angriff vielleicht entgehen zu können, weil das Backup – hoffentlich verschlüsselt – nicht mit demselben Angriffsvektor attackiert werden kann…
Aber natürlich gilt das skizzierte Problem nicht nur für Anwendungssoftware, sondern ebenso für die Betriebssysteme. Wenn also bspw. Ihre Drupal-Installation ebenso wie Ihr Backup aus Gründen der leichteren Administration auf demselben Betriebssystem bzw. dessen identischer Version verteilt abgelegt ist, wiederholt sich das Problem nur auf anderer Ebene.
Die Schlussfolgerung für eine Lösung kann also nur lauten, dass sowohl das Betriebssystem als auch die Anwendungssoftware eines Projektes sich im Kern unterscheiden sollten, *ohne* dass darunter die Benutzbarkeit leidet. Natürlich ist es im Prinzip egal, in welchem Datenbank-Management-System die Daten gespeichert und mit welcher Software sie abgerufen und dargestellt werden, so dass hier eine Diversifikation eigentlich nicht allzu schwierig sein sollte, selbst wenn man die Forderung erhebt und realisiert, dass diese Unterschiede für die Benutzer nicht sichtbar sein dürfen. Aber schon der Einsatz gängiger kommerzieller Software-Lösungen erlaubt dies i.d.R. nicht: Sei es, weil diese Software nur für ein Betriebssystem verfügbar ist oder weil für solche parallelen Mehrfachinstallationen natürlich auch mehrfache Lizenzen erworben werden müssen. Kostenlose, aber unfreie Software wie die PHP-Engine ZEND steht vor demselben Problem in nur leicht abgeschwächter Form, denn letztlich kommt es «auf die paar Euro» für die Lizenzen angesichts der Kosten für die Datenerhebung und ihre wissenschaftliche Auswertung nicht wirklich an – auch wenn die Forschungsförderinstitutionen sich sicherlich aktuell noch sträuben dürften, solche Kosten zu übernehmen: «Forschungsdatenmanagement» im 21. Jahrhundert…
Natürlich würde die Schaffung einer solchen verteilten und diversifizierten Dateninfrastruktur jedes einzelne Projekt ebenso überfordern wie die aktuellen Versuche, *langfristig* sicheres Datenmanagement durch die befristete Anstellung eines Mitarbeiters für «Forschungsdatenmanagement» aus kurzfristig bereitgestellten Mitteln des BMBF zu erreichen…
Auch hier bleibt m.E. nur der Schluss, dass sowohl die Entwicklung entsprechender digitaler Forschungsumgebungen als auch die Aushandlung globaler Verträge für wirklich verteilte Datenhaltung nur auf – mindestens – nationaler Ebene in einer gemeinsamen, *koordinierten* Kraftanstrengung durch eine Institution angegangen werden kann, die ähnlich wie Staatsbibliotheken, -museen und -archive einer gewissen «Ewigkeitsgarantie» unterliegen.
Dass sich auch hierfür bzw. aus diesem Grund und zur Erreichung dieses Ziels die Verwendung freier Software und freier Datenformate als absolute Notwendigkeit erweist, [5] verdeutlicht schon allein die Tatsache, dass gängige Projekte sicherlich *nicht* die finanziellen Ressourcen haben (werden), um in einer hoffentlich sehr langen Zukunft beliebig viele Lizenzen einer kommerziellen Softwarelösung

für eine einzige oder gar tatsächlich viele verteilte Datenhaltungen auf physisch getrennten Systemen zu erwerben. Und selbst *wenn* das Geld dafür vorhanden *wäre*, wäre es m.E. unverantwortlich, es eben genau *dafür* auch auszugeben und nicht für die chronisch unterfinanzierte Forschung selbst.

Um kurz in Erinnerung zu rufen, was alles zu beachten wäre, wenn man eine halbwegs sichere Lösung entwickeln wollte:

- Die Bediensoftware zum Zugriff auf die aufwendig erhobenen Daten und ihre interaktive Nutzung mit all ihren projektspezifischen Verknüpfungen und den diese widerspiegelnden Oberflächen müsste in verschiedenen Varianten vorliegen, die *nicht* durch Ausnutzung derselben Fehler gefährdet werden können.
- Dasselbe gilt für die darunter liegende Ebene der Datenbank-Management-Systeme und Programmierumgebungen sowie die – häufig ad hoc auszuführenden – Skriptsprachen.
- Erst recht gilt dasselbe für die eingesetzten Betriebssysteme und ihre Varianten/Versionen *und Compiler*.
- Ausschließliche Verwendung von Software, die nach heutigen Maßstäben als möglichst sicher angesehen werden kann, denn *jede* Software enthält Sicherheitslücken, die zu schliessen im Laufe immer längerer Speicherungs- und Nutzungsfristen immer schwieriger wird. (Doch wer benutzt schon PHP, oder…?)
- Aber selbst für die Hardware muss dies nach den Erfahrungen mit *Spectre* bzw. *Meltdown* inzwischen als gültig angesehen werden.
- Hinzu kommt die Forderung nach wirklich umfassender Dokumentation der «Logik» der für ein spezifisches Projekt angepassten Software: die gern so genannten «Suchmasken», die nur der sichtbare Ausdruck «eingebauter», den spezifischen Interessen und Anforderungen eines Forschungsprojekts entsprechender Logik(en) sind. Wenn Wissen im Verknüpfen von Informationen besteht – wovon ich fest überzeugt bin –, dann *müssen* diese Verknüpfungen ebenso Teil der Forschungsdaten und damit ihres Managements sein, wie die gesammelten oder erzeugten Daten selbst auch.

Das Problem der wirklich *langfristigen* Datensicherheit ist damit natürlich noch nicht ansatzweise gelöst! Aber immerhin sollte die Erfüllung dieser Anforderungen sie erleichtern.

Es dürfte klar sein, dass kein einzelnes Forschungsvorhaben – schon gar nicht in den historischen bzw. Geisteswissenschaften – diese Aufgaben auch nur im Ansatz lösen kann. Hier wären dieselben Forschungsförderungs-Institutionen und Ministerien in der Verantwortung, die von jedem einzelnen kleinen (oder großen) Projekt einen «Forschungsdaten-Management-Plan» verlangen, den sie selbst aktuell m.E. nicht befriedigend zu erstellen vermöchten. *Sie* müssten diese Infrastrukten zur Verfügung stellen – und ihre Benutzung für alle Projekte, die Steuergelder erhalten, ebenso verpflichtend machen wie den Open Access zu den Ergebnissen und Rohdaten…

5. LÖSUNGSANSÄTZE

Der oben bereits erwähnte Vint Cerf hat schon vor einigen Jahren einen Lösungsansatz vorgestellt, der auf den ersten Blick alle diese Probleme löst: das sog. *Digital Vellum* (Digitales Pergament), an dem sein Team nun bereits geraume Zeit arbeitet. [7] Allerdings ist diese Entwicklung längst noch nicht abgeschlossen. Und, soweit ich weiß, kann sie zwei zentrale Probleme nicht lösen:

- Beschränkungen durch Lizenzen, die bspw. die über lange Zeiträume beliebig häufige Installation einer verwendeten Software (idealerweise auch auf verschiedenen Betriebssystemen) verbieten.
- Die Vernetzung von Forschungsdatenbanken: Auch wenn es möglicherweise in den historischen und Geisteswissenschaften noch nicht üblich ist: Angesichts der umfassenden Verfügbarkeit von Internetanbindungen mit hohen Geschwindigkeiten (außer in Ländern der Dritten Welt oder in Deutschland außerhalb der Großstädte) dürfte es weit häufiger vorkommen, dass Bestandteile einer Datenbank, aus denen die Anzeige einer Webseite ad hoc erzeugt wird, von entfernten Servern geholt werden, wie dies bspw. bei der Facebook *Timeline* längst geschieht – und bei lästigen Werbeeinblendungen inzwischen seit Jahrzehnten üblich ist.

Selbst wenn solche technischen Möglichkeiten von den – im Vergleich zu kommerziellen Industrieanwendungen: – kleinen Projekten in den *Digital Humanities* noch nicht genutzt

werden, dürften dies über kurz oder lang schon aus ökonomischen Gründen notwendig sein: Es ist ja eigentlich einfach auch nicht nachzuvollziehen, warum bpsw. digitale Bilder mehrfach vorgehalten werden müssten, *ohne* dass es dabei überhaupt schon um den oben erwähnten Aspekt der verteilten Datenhaltung aus Sicherheitsgründen geht. Meistens dürfte die Begründung dafür in der Latenzzeit der Serverantwort liegen… oder in dem Problem der Unerreichbarkeit der originalen Daten bei deren Umzug auf ein anderes System mit anderer URL – obwohl solche Probleme in ein paar Zeilen Code zu lösen wären…

Dieses Problem über das Netz verteilter Daten, auf denen eine Anwendung basiert bzw. auf welche diese zugreift, ist m.E. auch im Ansatz des von mir hoch verehrten Alan Kay noch nicht mitgedacht: die verteilte Speicherung nicht nur der «Rohdaten» *und* der Beschreibung ihrer Beziehungen untereinander (z.B. in meist ja nicht wirklich gut menschenlesbarem XML), sondern eben auch der *gesamten* Software-Umgebung, die für ihre sofortige Nutzung notwendig ist. Es ist m.E. eine gefährliche Illusion zu glauben, die eigenen Daten würden den Forschern ferner Jahrzehnte oder gar Jahrhunderte schon interessant genug erscheinen, um sich dann ggf. die notwendige Software «drum herum» zu rekonstruieren. Die von Alan Kay vorgeschlagenen und prototypisch entwickelten digitalen «Cuneiform Tablets of 2015» [8] gehen deshalb auch davon aus, dass die Rekonstruktion der *gesamten* technischen Umgebung eines in ferner Zukunft aufgefundenen Datenspeichers (einer CD-ROM, DVD oder ähnlichen Trägers mit Daten *und* Software) anhand der aufgedruckten Anleitung nicht länger als ein «Nachmittagsprojekt» dauern darf. Bei den mir bekannten Forschungsprojekten ist eher das Gegenteil der Fall: Ohne wochen- oder gar monatelange Einarbeitung – nicht nur bezogen auf Datenstrukturierung und -kennzeichnungen, sondern erst recht auf die Software – ist bspw. deren Übernahme in ein neues Projekt oder gar «Wiederbelebung» längere Zeit nach Projektende und ohne Hilfe eines ursprünglich Mitwirkenden einfach fast nicht möglich.

6. SCHLUSS

Es ist klar, dass dies Maximalforderungen sind; aber erfahrungsgemäß werden diese im Zuge einer «Verhandlung» über das Wünsch- und technisch Machbare ohnehin – wie in jedem Handel auf dem Basar reduziert werden müssen: Erinnert sich noch jemand an «The Cathedral and the Bazaar»? [9] Um im Bild zu bleiben: Zur Zeit werden m.E. in Forschungsprojekten der historischen und Geisteswissenschaften weit überwiegend kleine, aber sehr idiosynkratische «Kathedralen» errichtet – und zwar auf dem «Sand» nur kurzfristig überlebensfähiger Software und mit «Mörtel», der mit Ablaufen seiner Lizenz oder beim nächsten entdeckten Sicherheitsproblem zu «Butter» wird. Dann nützen auch die besten «Steine» nichts mehr. Um das zu verstehen, muss man wohl sicher kein Statiker, Architekt oder auch nur Architekturhistoriker sein. Aus meiner Sicht lassen sich allerdings gar keine verhandelbaren Abstriche an diesen «Maximalforderungen» machen, denn es gibt m.W. keine Lösungen, die mit irgendwelchen Abstrichen das Geforderte leisten könnten. Ich lasse mich aber gern eines Besseren belehren!

7. LITERATURHINWEISE

[1] Kulawik, Bernd: Digitales Kuratieren – und dann? – In: Staatliche Museen zu Berlin (Hg.): EVA Berlin 2016, S. 75–82 [PDF]

[2] Kulawik, Bernd: Wie man das Verschwinden unserer Daten im «Digitalen Schwarzen Loch» und somit ein «dunkles Informationszeitalter» vermeiden könnte. – In: Staatliche Museen zu Berlin (Hg.): EVA Berlin 2017, S. 203–210. [PDF]

[3] Kulawik, Bernd: «If there are documents you really care about: Print them out!» (after Vint Cerf, 2015). – In: Loizides, F.; Schmidt, B. (Hg.): Positioning and Power in Academic Publishing = Proceedings of the 20th International Converence on Electronic Publishing, Göttingen, 2016, S. 23–27 [PDF]

[4] Sample, Ian: Google boss warns of `forgotten century' with email and photos at risk. – The Guardian, Friday 13 Feb 2015. www.theguardian.com/technology/2015/feb/13

[5] Kulawik, Bernd: Why and how to avoid complex non-free software in Digital Humanities projects. – In: *Information Services & Use* 36 (2016), S. 203–210 [PDF]

[6] Kulawik, Bernd: From Top-Down to Network: Long-Time Perspectives of Scientific Publication: www.kunstgeschichte-ejournal.net/165

[7] Vgl. z.B. die Präsentation http://wirth-symposium. ethz.ch/slides/cerf.pdf von 2014.

[8] Kay, Alan; Nguyen, Long Tien: The Cuneiform Tablets of 2015. – Los Angeles: 2015 vpri.org/pdf/tr2015004_cuneiform.pdf

[9] Raymond, Eric S.: The Cathedral and the Bazaar. – O'Reilly: 1999: www.catb.org